AD-A269 536

Universit
of Southern
California

NATURAL LANGUAGE PROCESSING BY THE PENMAN PROJECT AT USC/ISI

Eduard H. Hovy
USc/Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292

June 1993
ISI/RR-93-353

INFORMATION SCIENCES INSTITUTE

310/822-1511

4676 Admiralty Way/Marina del Rey/California 90292-6695

12

NATURAL LANGUAGE PROCESSING BY THE PENMAN PROJECT AT USC/ISI

Eduard H. Hovy
USc/Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292

June 1993
ISI/RR-93-353

REPORT DOCUMENTATION PAGE

FORM APPROVED
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
June 1993

3. REPORT TYPE AND DATES COVERED
Research Report

4. TITLE AND SUBTITLE
Natural Language Processing By the Penman Project

5. FUNDING NUMBERS
MDA903-87-C-641

6. AUTHOR(S)
Eduard Hovy

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
USC INFORMATION SCIENCES INSTITUTE
4676 ADMIRALTY WAY
MARINA DEL REY, CA 90292-6695

8. PERFORMING ORGANIZATON REPORT NUMBER
RR-353

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
ARPA
3701 Fairfax Drive
Arlington, VA 22203

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12A. DISTRIBUTION/AVAILABILITY STATEMENT
UNCLASSIFIED/UNLIMITED

12B. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The Penman project at USC/ISI has been conducting research in computational Natural Language Processing since 1978, mainly in the area of language generation. This research includes work on single-sentence realization as well as multi-sentence text planning for descriptions and explanations. Over the past few years, the project's focus has broadened to include research on Machine Translation, including parsing and the semi-automated construction of large semantic knowledge bases and lexicons of various languages, as well as research on the automated planning of multimedia and multimodal communications in general. This paper provides an overview of the different research directions.

14. SUBJECT TERMS
Natural Language Process, computational linguistics, Information Sciences Institute of USC, Penman, generation, parsing, text planning, computational studies of discourse

15. NUMBER OF PAGES
12

16. PRICE CODE

17. SECURITY CLASSIFICTION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reoprts. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to stay within the lines to meet optical scanning requirements.

Block 1. Agency Use Only (Leave blank).

Block 2. Report Date. Full publication date including day, month,a nd year, if available (e.g. 1 jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken fr~ the part of the report that provides the m meaningful and complete information. Wl report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element numbers(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	PR	- Project
G	- Grant	TA	- Task
PE	- Program Element	WU	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the repor.

Block 9. Sponsoring/Monitoring Agency Names(s) and Address(es). Self-explanatory

Block 10. Sponsoring/Monitoring Agency Report Number. (If known)

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of ...; To be published in... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD	- See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE	- See authorities.
NASA	- See Handbook NHB 2200.2.
NTIS	- Leave blank.

Block 12b. Distribution Code.

DOD	- Leave blank.
DOE	- Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA	- Leave blank.
NTIS	- Leave blank.

Block 13. Abstract. Include a brief (Maximum 200 words) factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code (NTIS only).

Blocks 17.-19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contins classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

NATURAL LANGUAGE PROCESSING BY THE PENMAN PROJECT AT USC/ISI

Eduard H. Hovy

Information Sciences Institute
of the University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
U.S.A.
tel: 310-822-1511
fax: 310-823-6714
email: HOVY@ISI.EDU

Abstract

The Penman project at USC/ISI has been conducting research in computational Natural Language Processing since 1978, mainly in the area of language generation. This research includes work on single-sentence realization as well as multi-sentence text planning for descriptions and explanations. Over the past few years, the project's focus has broadened to include research on Machine Translation, including parsing and the semi-automated construction of large semantic knowledge bases and lexicons of various languages, as well as research on the automated planning of multimedia and multimodal communications in general. This paper provides an overview of the different research directions.

1 Overview

Currently, Natural Language Processing (NLP) work in the Penman project at USC/ISI is organized around five principal theoretical efforts within the general area of Machine Translation:

1. Natural language generation (single-sentence realization).
2. Discourse structure development (paragraph-length text planning).
3. Knowledge resource acquisition and management (semi-automated semantic knowledge base construction and multilingual lexicon acquisition).

4. Natural language understanding (single-sentence parsing).
5. Multimedia and multimodal communication (presentation planning and dynamic information-to-medium allocation).

USC/ISI is a non-profit organization of about 200 people conducting research into various aspects of Computer Science. The Penman project is part of the Intelligent Systems Division, whose members are investigating a number of questions in the general area of Artificial Intelligence (AI). Other projects in this division include:

- **Loom**: Knowledge representation in the KL-ONE framework
- **SIMS**: Single integrated access to numerous databases
- **EXPECT/EES**: Explainable expert systems
- **SOAR**: General architecture for intelligent reasoning
- **DRAMA**: Software development environment management systems
- **Humanoid**: Multimedia interface construction environment

2 Pangloss

The PANGLOSS Machine Translation (MT) project is a collaborative effort between USC/ISI, the Center for Machine Translation (CMT) at Carnegie Mellon University, and the Computing Research Laboratory (CRL) at New Mexico State University. Most of the current research in the Penman project is directed by the needs and requirements of the PANGLOSS system.

PANGLOSS is a human-assisted MT system with the following features:

- Initial languages are Spanish to English. Japanese as input language is being added starting mid-1993. Additional possible input and output languages are German and Chinese.
- The initial application domain is newspaper texts on financial Merger and Acquisition transactions.
- Human assistance can occur (via a program called the Augmentor) during the translation. When a process module runs into trouble it calls the Augmentor and then through various manipulations the user helps it, or acquires new information such as lexical items.
- System development is phased, with increasing Automation (that is, the application domain is kept constant and the output quality as well.) Initially, PANGLOSS was principally a human aid, an editing tool with lexicons and dictionaries and word processors. As more capabilities are added, the human operator does less, with the aim of minimizing human intervention by the end of 1995.

- The system uses an Interlingua as internal representation of the input text. Interlingua terms are defined in an extensive taxonomy of approximately 50,000 concepts called the Ontology.
- For Spanish parsing, CRL's parser ULTRA and Spanish grammar [Farwell & Wilks 91] are used. ULTRA's output contains a mixture of syntactic and semantic information, following the theory of preference semantics. CRL is also responsible for the creation of the Spanish lexicon and the collection of other useful textual resources. ULTRA is written in Quintus Prolog.
- For the semantic analysis of both Spanish and Japanese, and for the construction of the Interlingua statement corresponding to the input, CMT is responsible. CMT is also responsible for the system architecture, the operator interface (including the Augmentor, WordPerfect and emacs text editing tools, etc.) [Frederking et al. 93], and for the definition of the Interlingua notation. All the CMT software is written in CMU Common Lisp.
- For generation, USC/ISI's Penman system [Penman 88, Matthiessen & Bateman 91] is used in tandem with its sentence planning software. The Penman system follows the theory of Systemic Functional Linguistics [Halliday 85]. USC/ISI is also responsible for the creation of the English lexicon and the creation of the concept Ontology, as well as for the development of the Japanese parser. All the software is written in Lisp.

3 Single-Sentence Natural Language Generation (Penman)

PENMAN is a natural language sentence generation program developed at USC/ISI since 1982. It provides computational technology for generating English sentences, starting with input specifications of a non-linguistic kind. The culmination of a continuous research effort since 1978, Penman embodies one of the most comprehensive computational generators of English sentences in the world.

Three research goals underlie Penman: to provide a framework in which to conduct investigations into the nature of language, to provide a useful and theoretically motivated computational resource for other research and development groups and the computational community at large, and eventually to provide a text generation system that can be used routinely by computer system developers.

Penman consists of a number of components. Nigel, the English grammar, is the heart of the system. Based on the theory of Systemic Functional Linguistics (a theory of language and communication developed by Halliday and others [Halliday 85, Halliday 73, Halliday 66], and used in various other AI applications, such as in SHRDLU [Winograd 72]), Nigel is a network of over 700 nodes called *systems*, each node representing a single minimal grammatical alternation. In order to generate a sentence, Penman traverses the network guided by its inputs and default settings. At each system node, Penman selects a feature until it has assembled enough features to fully specify a sentence. After constructing a syntax tree and choosing words to satisfy the features selected, Penman then generates the English sentence.

The Nigel grammar is described in, among others, [Mann & Matthiessen 83, Matthiessen 84]. In order for grammarians to use or extend Nigel, they need simply load it on a computer; Nigel's window interface is tailored to support research on grammar construction and control.

Besides Nigel, Penman also contains a number of information resources, such as a lexicon of 50,000+ English words (containing word definitions, inflectional forms, etc.) and the Penman Upper Model, a very general taxonomic model of the world [Bateman et al. 89]. This taxonomy acts to link the terms in a user's application domain to the terms used within Penman. It is based on the distinctions made in English — for example, since objects are treated differently in English than actions, actions and objects are placed in different classes in the model — and is represented as a generalization hierarchy with property inheritance. In order to use Penman, a user must define a lexicon of domain-specific words and also provide a model of domain-specific entities which is then linked to the Upper Model. Penman includes a lexical acquisition tool, LAPITUP, that allows a person with relatively little training to create lexical items for Penman's use. The structure of Penman is described in detail in [Mann 82, Matthiessen & Bateman 91]. Its use is described in the Penman documentation [Penman 88].

Penman is designed to be used effectively by people with various degrees of linguistic and computational sophistication. Depending on their interests, different people will use different parts of it, feed it different types of inputs, and expect different types of outputs. A systemic linguist would interact mainly with Nigel, controlling selections within systems, and studying the resulting output feature collections and realizations. A computational linguist would interact with the whole system, providing semantic specifications of the sentences desired after having built a lexicon and a model of the domain of discourse. A computer scientist would use Penman purely as an output module to convert the output of some program into English, and after defining a lexicon and domain model, would use as many of Penman's internal input building functions as possible.

At USC/ISI, Penman is currently being used primarily as the output generator of the PANGLOSS project.

The Penman sentence generator is written in Common Lisp and currently operates on Sun SPARCStations, Sun 4s, TI Explorer and Symbolics Lisp machines, and Macintosh-II computers (with 8 mb or more memory). Penman has been distributed to over 90 sites worldwide, and has been used for graduate-level instructional purposes at various universities, as well as forming part of several Ph.D. dissertation efforts. On the Mac, the full system occupies about 7.5 megabytes and generates a two-clause sentence in about 20 seconds; on a TI Explorer, it generates the same sentence in under 2 seconds. For further information on Penman please contact the author.

4 Discourse Structure Development (Text Planning)

Over the last several years, members of the project have been investigating the internal structure of discourse and the computational planning and generation of coherent multisentential paragraphs. A theory of the interclausal relationships that govern discourse structure,

called Rhetorical Structure Theory (RST) [Mann & Thompson 85, Mann & Thompson 88a, Matthiessen & Thompson 88], was developed after extensive analysis of hundreds of texts of various genres. The analysis concluded that English text is coherent by virtue of so-called rhetorical relations that hold between clauses and blocks of clauses, and identified about 25 basic relations for English. These relations, such as SEQUENCE, PURPOSE, and ELABORATION are usually identified by key words or phrases (such as "then", "in order to", and "e.g.", respectively).

In order to plan multisentence paragraphs by computer, one requires both a sound theory of text organization and an algorithm that can make efficient use of it. The theory is provided by RST; the algorithm by an adaptation of the top-down hierarchical expansion planning system NOAH (see [Sacerdoti 75]). A series of text structure planners have been developed by members of the Penman project and visitors to plan coherent paragraphs which achieve communicative goals of affecting the hearer's knowledge in some way. The planners operate in conjunction with some application program (such as a database access system or expert system) and employ Penman to generate the individual sentences. From the application program, the planners accept one or more communicative goals, as well as in some cases a set of clause-sized input entities that represent the material to be generated. Using operationalized RST relations and other text plans, they construct a tree that embodies the paragraph structure, in which nonterminal nodes are RST relations and terminal nodes contain the material to be communicated. This text planning process was initially developed in [Hovy 88, Hovy 90a], and has been greatly extended by several other projects, both at USC/ISI and elsewhere. A general overview of this work appears in [Hovy 93].

One major extension involves the number of interclausal discourse structure relations. In one study, the author collected and taxonomized over 300 relations from a variety of sources; this collection was then further elaborated and reorganized. For a fairly extensive description see [Hovy & Maier 93].

A second extension performed at USC/ISI is the automated planning of certain types of text formatting. In [Hovy & Arens 91], the communicative semantics of certain text formatting devices (such as enumerated lists, itemizations, footnotes, appendices, etc.) is described in terms of RST relations, and the automated planning of formatted paragraphs of text is illustrated.

In separate work, members of the EES/EXPECT project at USC/ISI built the EES text planner along the same lines as the initial Penman text structurer, incorporating a greatly expanded text plan library using a notation oriented toward intentionality [Moore & Swartout 88, Paris 90, Moore 89, Moore & Paris 89]. This planner's text plan contains the intentional, attentional, and rhetorical structures of the explanations it generates for EES expert systems. By recording the goal structure of the text being produced, the rhetorical strategies employed, and any assumptions made about the user's goals and knowledge, the EES planner is able to reason about previous responses in order to interpret a user's follow-up questions in the ongoing conversation and determine how to clarify a response when necessary. Furthermore, by having multiple explanation strategies, the system is able to select the one that is most appropriate for a specific user, and to choose an alternate strategy to recover from failure.

In later work, a new text planner that combines some of the ideas of the Penman and the

EES planners has been developed, primarily by a visiting graduate students from Germany, Elisabeth Maier. This planner is described in [Hovy et al. 92, Maier 93]. Although none of the text planners are relevant to the PANGLOSS system, certain aspects of text planning, including the determination of sentence length, clause aggregation to remove redundancies, and some types of lexical choice, are. These text planning tasks are being incorporated into the PANGLOSS Sentence Planner, which converts the PANGLOSS internal interlingual notation into the Penman input format.

5 Knowledge Resource Acquisition and Management

This research direction addresses the need for acquiring large semantic and lexical knowledge resources, both for Penman-specific work and to support the sharing of knowledge across PANGLOSS modules at other sites. Since PANGLOSS uses an Interlingua, which by definition is language-neutral, an obvious candidate for shared knowledge is the definitional framework of the Interlingua symbols. This is the point of least representational difference (lexical, syntactic, etc.) between parsers, analyzers, and generator.

The PANGLOSS Ontology is a taxonomy of approximately 50,000 symbols that represent the semantic meanings conveyed in translations. The Ontology is being constructed at USC/ISI by Dr. Kevin Knight, by extracting knowledge from a variety of sources. It is represented in Loom, FrameKit, and Prolog, and is distributed with appropriate access routines to the other PANGLOSS sites.

The topmost levels of the Ontology, which we call the Ontology Base (OB), consist of approx. 400 terms. The OB contains nodes that represent generalized distinctions required for the processing of the parsers, analyzers, and generator. While the idiosyncratic processing requirements of each lexeme are stored either in a lexicon (for morphological and syntactic information) or in the Ontology body (for semantic information), general semantic and syntactic patterns are captured as nodes in the OB. The OB is a merge of the Penman Upper Model (based on Systemic-Functional Linguistics), the top-level ONTOS ontology (a semantic network; see [Nirenburg & Defrise 92]), and, for nouns, the LDOCE semantic categories. It maintains the distinctions present in the Upper Model so that all subordinated Ontology terms can be properly generated in English; it maintains the LDOCE categories so that ULTRA can make the necessary distinctions when parsing nouns; and it maintains the ONTOS distinctions so that semantic analysis can proceed properly. The function of the Ontology Base and its relation with the Interlingua are described in [Hovy & Nirenburg 92].

The primary two sources for the Ontology body are the Longman Dictionary of Contemporary English (LDOCE) [LDOCE 78] and the semantic database WordNet [Miller 85]. LDOCE senses are tagged with useful syntactic, semantic, and pragmatic information that can be extracted automatically. However, since LDOCE senses are not grouped by synonymy and are not arranged in a deep hierarchy, the taxonomization of WordNet served as an initial basis of construction. To construct the main body of the Ontology, work was performed to automatically merge LDOCE and WordNet by discovering pairs of corresponding senses [Knight 93].

In addition to housing the symbols to represent semantic meaning, the Ontology contains pointers from each symbol to appropriate lexical items in various languages. The Penman lexicon currently contains about 50,000 spelling forms (corresponding to approx. 70,000 words). Though lexicons of similar size of Japanese, Chinese, and Spanish have been acquired, their items have at the time of writing not yet been formatted in the generic lexicon form or linked to the Ontology.

6 Natural Language Understanding (Parsing)

Early research in parsing in the Penman project involved the construction of a prototype parser to use the Nigel grammar, enabling its bidirectional use for both language generation and understanding. Based on a widely used unification-based parsing system developed at SRI (PATR-II [Shieber 84]), the prototype parser used a form of Nigel, rewritten in the notation of Functional Unification Grammar (FUG) [Kay 85], to accommodate a fuller range of grammatical descriptions, including descriptions containing disjunctive and conditional information (see [Kasper 88a, Kasper 88b]).

The prototype parser operated using methods of unification, which is why it required the rewritten form of the grammar in FUG. However, recent advances in the theory of representation languages make possible the representation of the grammar in Loom instead of in FUG. This approach enables a new integrated treatment of syntax and semantics, using Loom's subsumptive classifier instead of the unifier. The method, which is currently being implemented, is a novel parsing technique and holds great promise: not only is the parsing process likely to be much simpler than traditional parsing (in which syntactic and semantic parsing proceed under different mechanisms and have to be linked explicitly), but it also makes use of the functionally oriented Systemic grammar Nigel, which is one of the larger computational grammars of English, and because of the flexibility of its system network notation is rather amenable to the parsing of semantic, thematic, and other information.

Once work was completed at USC/ISI to incorporate the ability to perform inference over disjunctions in Loom, syntactic and semantic knowledge could be represented in the same knowledge representation system, and parsing could be performed with respect to them both simultaneously. The prototype parser accesses semantic and syntactic information as soon as it is relevant in a straightforward and direct fashion using a single mechanism, the Loom classifier, for its primary inferencing operation. The potential benefits of an integrated, single-operation parsing approach are manifest: simplification of process, reduction of processing overhead, and facilitation of representation of dependencies between syntax and semantics.

In a completely separate development, plans are underway for the construction of a Japanese parser at USC/ISI for use in the PANGLOSS project. The construction of this parser will employ statistical techniques to ensure robustness and wide coverage of the application domains as well as symbolic techniques to ensure the depth of the parsed results. This work is scheduled to begin in late 1993.

7 Multimedia And Multimodal Communication

Although no active funding or formal project has existed yet, members of the Penman project have for several years performed some research on several core issues in automated multimedia presentation planning. Usually, the work involved one or more graduate students who visited USC/ISI to complete their Master's theses.

One of the core issues involves the generalization of techniques for the automated planning of texts to apply also to multimedia presentations. A second area addresses the central question of information-to-medium allocation: which information should be apportioned to which display medium? In an ongoing study, characteristics of information, media, and modalities are being analyzed and a dynamic allocation algorithm is being developed. Some overall theoretical ideas are summarized in [Arens et al. 93a, Hovy & Arens 90] and one of the prototype systems constructed is described in [Arens et al. 93b].

8 Other Interests

In addition to the work described above, ISI researchers, in some cases in collaboration with other researchers, have pursued or plan to pursue work on the following questions:

- **Register-Controlled Generation of Variations:** The definition and use of register in order to determine the selection and organization of material, constituent head, and lexical entity, in order to tailor the generated text to the level of sophistication of the reader. Drs. John Bateman and Cécile Paris from USC/ISI. See [Bateman & Paris 89a, Bateman & Paris 89b].
- **Semantic Information Retrieval:** The use of the Ontology as an overarching index structure under which to index a library of texts and pictures, enabling the multilingual access of appropriate objects through the use of the lexicons attached to the Ontology. Drs. Eduard Hovy and Kevin Knight from USC/ISI, in collaboration with Dr. Hatte Blejer from SRA Corporation, Washington, DC.
- **Speech Generation:** The addition into the grammar of features to control the realization of intonational contours in order to achieve desired communicative effects. Dr. John Bateman with Prof. Bea Oshika from the Portland State University, OR.

9 Collaborations

In order to promote increased development of various computational aspects of Systemic Linguistics, the project partakes in a multinational collaboration, in which various partners have different focuses of research, but which are all oriented around some aspect of Penman. All work is shared among all the partners and periodic updates ensure that everyone uses the same basic mechanisms in their investigations. This collaboration started in September 1989. The partners are:

- The Penman project at USC/ISI, Marina del Rey, USA. Roughly speaking, USC/ISI acts as a clearing-house for the computational implementation and distribution of Penman and other software, while supporting various aspects of research. Contact persons:

  Dr. Eduard Hovy, Dr. Kevin Knight
  email: HOVY@ISI.EDU, KNIGHT@ISI.EDU

- Members of the Linguistics Department of the University of Sydney, Australia. The Linguistics Department group in Sydney pursues fundamental work on grammar development, Japanese and Chinese grammars for Penman, and parsing. Contact person:

  Prof. Christian Matthiessen
  email: XIAN@BRUTUS.EE.SU.OZ.AU

- The KOMET project at IPSI, Darmstadt, Germany. IPSI supports research on generation of German, a German Upper Model, text planning, and lexical choice. Contact person:

  Dr. John Bateman
  email: BATEMAN@DARMSTADT.GMD.DE

10 Natural Language Researchers and Publications

At the time of writing, the Penman project consists of Drs. John Bateman (part-time), Eduard Hovy (project leader), Kevin Knight, and Mr. Richard Whitney. It has three open positions. In addition, several visitors are usually working at USC/ISI at any point.

Other projects with associated research include the EES/EXPECT project (Dr. Cécile Paris and Mr. Vibhu Mittal), the SIMS project (Dr. Yigal Arens), the IDOC project (Dr. Lewis Johnson), and the Division Director, Dr. William Swartout.

A number of people have worked as project members or consultants in the past; the list includes Drs. Ken Church, Susanna Cumming, Cecilia Ford, Peter Fries, Michael Halliday, Robert Kasper, Christian Matthiessen, Johanna Moore, Norman Sondheimer, Sandra Thompson; Ms. Lynn Poulton; and Messrs. Robert Albano, Thomas Galloway, and Mick O'Donnell. In addition, for many years the Penman project has benefitted from the work of visiting researchers too numerous to list.

The group embodies a combination of Computer Science and Linguistics (in earlier years the proportion was about 70% Computer Science and 30% Linguistics). We maintain active interaction with linguists who serve as consultants, primarily in the areas of discourse, grammar, lexical knowledge and speech processing. We also maintain contact with academic departments of several universities in the U.S. and abroad, and regularly employ graduate students from USC, UCLA, and other institutions.

The group has an active publication record; a list of technical reports can be sent on request to Ms. Kary Lau (email: KARY@ISI.EDU).

11 Conclusion

The Penman project is always in search of new opportunities for growth and new collaborations. The group has hosted a number of shorter-term visitors and Fulbright scholars, and attempts to foster an open, friendly, and positive research environment. For further information, please contact the author.

References

[Arens et al. 93a] Arens, Y., Hovy, E.H., and Vossers, M. 1993. Describing the Presentational Knowledge Underlying Multimedia Instruction Manuals. In *Intelligent Multimedia Interfaces*, M. Maybury (ed).

[Arens et al. 93b] Arens, Y., Hovy, E.H., and Van Mulken, S. 1993. Structure and Rules in Automated Multimedia Presentation Planning. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI*, Chambery.

[Bateman et al. 89] Bateman, J.A., Kasper, R.T., Moore, J.D., and Whitney, R.A. 1989. A General Organization of Knowledge for Natural Language Processing: The Penman Upper Model. Unpublished research report, USC/Information Sciences Institute, Marina del Rey.

[Bateman & Paris 89a] Bateman, J.A. and Paris, C.L. 1989. Phrasing a Text in Terms the User can Understand. *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI*, Detroit, MI.

[Bateman & Paris 89b] Constraining the Deployment of Lexicogrammatical Resources during Text Generation: Toward a Computational Instantiation of Register Theory. 1989. Presented at the *Sixteenth International Systemics Workshop*, Helsinki.

[Farwell & Wilks 91] Farwell, D. and Wilks, Y. 1991. ULTRA: A Multilingual Machine Translator. *Proceedings of the Third MT Summit*, Washington, DC (19–24).

[Frederking et al. 93] Frederking, R., Grannes, D., Cousseau, P. and Nirenburg, S. An MAT Tool and its Effectiveness. 1993. *Proceedings of the 1993 ARPA Human Language Workshop*.

[Halliday 66] Halliday, M.A.K. 1966. Some Notes on 'Deep' Grammar. *Journal of Linguistics* 2(1) (57–67).

[Halliday 73] Halliday, M.A.K. 1973. *Explorations in the Functions of Language*. Edward Arnold: London.

[Halliday 85] Halliday, M.A.K. 1985. *Introduction to Functional Grammar*. Edward Arnold Press: London.

[Hovy 88] Hovy, E.H. 1988. Planning Coherent Multisentential Text. *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, Buffalo.

[Hovy 90a] Hovy, E.H. 1990. Approaches to the Planning of Coherent Text. In *Natural Language in Artificial Intelligence and Computational Linguistics*, Paris, C.L., Swartout, W.R. and Mann, W.C. (eds). Boston: Kluwer (83–102).

[Hovy 93] Hovy, E.H. 1990. Automated Discourse Generation using Discourse Structure Relations. *Artificial Intelligence* (Special Issue on Natural Language Processing), to appear.

[Hovy & Arens 90] . Hovy, E.H. and Arens, Y. 1990. When is a Picture Worth a Thousand Words? — Allocation of Modalities in Multimedia Communication. Presented at the *AAAI Symposium on Human-Computer Interfaces*, Stanford.

[Hovy & Arens 91] Hovy, E.H. and Arens, Y. 1991. Automatic Generation of Formatted Text. *Proceedings of the 8th AAAI Conference*, Anaheim.

[Hovy & Maier 93] Hovy, E.H. and Maier, E. 1993. Parsimonious and Profligate: How Many and Which Discourse Structure Relations? *Discourse Processes*, to appear.

[Hovy et al. 92] Hovy, E.H., Lavid, J., Maier, E., Mittal, V., and Paris, C.L. 1992. Employing Knowledge Resources in a New Text Planner Architecture. In *Aspects of Automated Natural Language Generation*, R. Dale, E. Hovy, D. Rösner, O. Stock (eds). Heidelberg: Springer Verlag Lecture Notes in AI number 587 (57–72).

[Hovy & Nirenburg 92] Hovy, E.H. and Nirenburg, S. 1992. Approximating an Interlingua in a Principled Way. *Proceedings of the 1992 ARPA Speech and Natural Language Workshop*, Arden House.

[Kasper 88a] Kasper, R.T. 1988. Systemic Grammar and Functional Unification Grammar. In *Systemic Functional Approaches to Discourse*, Benson J. and Greaves W. (eds). Ablex: Norwood.

[Kasper 88b] Kasper, R.T. 1988. An Experimental Parser for Systemic Grammars. In *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest.

[Kay 85] Kay, M. 1985. Parsing in Functional Unification Grammar. In *Natural Language Parsing*, Dowty, D., Karttunen, L. and Zwicky, A. (eds). Cambridge University Press: Cambridge.

[Knight 93] Knight, K. 1993. Building a Large Ontology for Machine Translation. *Proceedings of the 1992 ARPA Human Language Workshop*, Princeton.

[LDOCE 78] *Longman Dictionary of Contemporary English*. 1978. Essex.

[Maier 93] Maier, E. Ph.D. dissertation, forthcoming.

[Mann 82] Mann, W.C. 1982. The Anatomy of a Systemic Choice. USC/ISI Technical Report RR-82-104.

[Mann & Matthiessen 83] Mann, W.C. and Matthiessen, C.M.I.M. 1985. Nigel: A Systemic Grammar for Text Generation. In *Systemic Perspectives on Discourse: Selected Papers Papers from the Ninth International Systemics Workshop*, Benson, R. and Greaves, J. (eds), Ablex: London. Also available as USC/ISI Research Report RR-83-105.

[Mann & Thompson 85] Mann, W.C. and Thompson, S.A. 1985. Assertions from Discourse Structure. In *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society*, Berkeley. Also available as USC/ISI Research Report RS-85-155.

[Mann & Thompson 88a] Mann, W.C. and Thompson, S.A. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3) (243–281).

[Matthiessen 84] Matthiessen, C.M.I.M. 1984. Systemic Grammar in Computation: The Nigel Case. In *Proceedings of 1st Conference of the European Association for Computational Linguistics*, Pisa. Also available as USC/ISI Research Report RR-84-121, 1984.

[Matthiessen & Bateman 91] Matthiessen, C.M.I.M. and Bateman, J.A. 1991. *Systemic-Functional Linguistics in Language Generation: Penman.*

[Matthiessen & Thompson 88] Matthiessen, C.M.I.M. and Thompson, S. 1988. The Structure of Discourse and Subordination. In *Clause Combining*, Haiman, J. and Thompson, S. (eds), John Benjamins: Amsterdam.

[Miller 85] Miller, G.A. 1985. WordNet: A Dictionary Browser. *Information in Data: Proceedings of the 1st Conference of the UW Centre for the New Oxford Dictionary*. University of Waterloo, Canada.

[Moore 89] Moore, J.D. 1989. *A Reactive Approach to Explanation in Expert and Advice-Giving Systems*. Ph.D. dissertation, University of California (Los Angeles).

[Moore & Paris 88] Moore, J.D. and Paris, C.L. 1988. Constructing Coherent Texts Using Rhetorical Relations. In *Proceedings of the 10th Cognitive Science Society Conference*, Montreal.

[Moore & Paris 89] Moore, J.D. and Paris, C.L. 1989. Planning Text for Advisory Dialogues. *Proceedings of the 27th ACL Conference*, Vancouver (67–75).

[Moore & Swartout 88] Moore, J.D. and Swartout, W.R. 1990. A Reactive Approach to Explanation. In Paris, C.L., Swartout, W.R. and Mann, W.C. (eds), *Natural Language in Artificial Intelligence and Computational Linguistics*, Kluwer.

[Nirenburg & Defrise 92] Nirenburg, S. and Defrise, C. 1992. Application-Oriented Computational Semantics. In R. Johnson and M. Rosner (eds.), *Computational Linguistics and Formal Semantics*. Cambridge: Cambridge University Press.

[Paris 90] Paris, C.L. 1988. Generation and explanation: Building an explanation facility for the Explainable Expert Systems framework. In Paris, C.L., Swartout, W.R. and Mann, W.C. (eds), *Natural Language in Artificial Intelligence and Computational Linguistics*, Kluwer. 1990.

[Penman 88] *The Penman Primer, User Guide, and Reference Manual*. 1988. Unpublished USC/ISI documentation.

[Sacerdoti 75] Sacerdoti, E. 1975. *A Structure for Plans and Behavior*. North-Holland: Amsterdam.

[Shieber 84] Shieber, S.M. 1984. The Design of a Computer Language for Linguistic Information. In *Proceedings of the 10th International Conference on Computational Linguistics*, Stanford.

[Winograd 72] Winograd, T. 1972. *Understanding Natural Language*. Academic Press: New York.